VERİ MADENCİLİĞİNE GİRİŞ

Veri Madenciliği Giriş

- İçinde yaşadığımız bilişim çağında elektronik ortamda mevcut verinin hızlı artışı ve bilginin fazlalaşması sebebiyle öncelikle, genelde Veri Tabanlarında Bilgi Keşfi olarak adlandırılan yeni bir paradigma ortaya çıkmıştır. Daha yaygın bir kullanımla bu alana Veri Madenciliği denilmektedir.

Veri Madenciliği Tanımları (1/2)

- Veri Madenciliği(Data Mining): Büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların aranmasıdır. (Knowledge Discovery in Databases)
- Daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veritabanlarından elde edilmesi ve bu bilgilerin işletme kararları verilirken kullanılmasıdır.
- Büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme işidir.
- Yapısal veritabanlarında depolanmış verilerden geçerli, yeni, potansiyel olarak yararlı ve nihayetinde anlaşılabilir örüntülerin tanımlanması işlemidir.

Veri Madenciliğinin Tarihçesi (1/4)

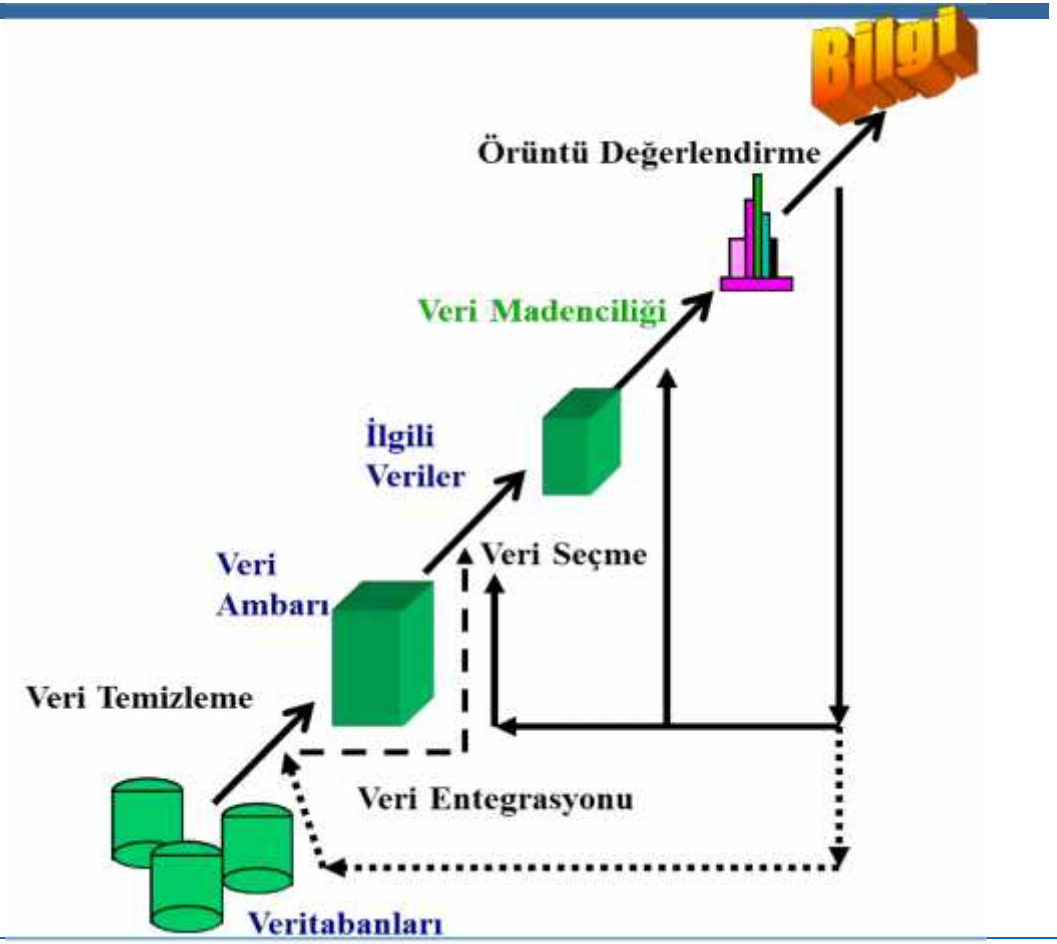
- Data FishingData Dredging: 1960
  - istatistikçiler
- Data Mining: 1990
  - veritabanı kullanıcıları, ticari
- Knowledge Discovery in Databases (KDD): 1989
  - Yapay zeka, makine öğrenmesi toplulukları
- Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction,...

Bilgi Keşfi

- Teoride veri madenciliği bilgi keşfi işleminin aşamalarından biridir.
- Pratikte veri madenciliği ve bilgi keşfi eş anlamlı olarak kullanılır.
- Veri madenciliği teknikleri veriyi belli bir modele uydurur.
  - veri içindeki örüntüleri bulur
  - örüntü: veri içindeki herhangi bir yapı
- Sorgulama ya da basit istatistik yöntemler veri madenciliği değildir.

- Büyük veri kaynaklarından yararlı ve ilginç bilgiyi bulmak
- Bulunan bilgi
  - gizli,
  - önemli,
  - önceden bilinmeyen,
  - yararlı olmalı.

Bilgi Keşfi


Bilgi
Örüntü Değerlendirme
Veri Madenciliği
İlgili
Veriler
Veri Seçme
Veri
Ambarı
Veri Temizleme
Veri Entegrasyonu
Veritabanları

Bilgi Keşfinin Aşamaları

- Veri Temizleme : Gürültülü ve tutarsız verileri çıkarmak
- Veri Bütünleştirme: Birçok data kaynağını birleştirebilmek
- Veri Seçme : Yapılacak olan analiz ile ilgili olan verileri belirlemek
- Veri Dönüşümü : Verinin veri madenciliği yöntemine göre hale dönüşümünü gerçekleştirmek
- Veri Madenciliği : Verilerdeki örüntülerin belirlenmesi için veri madenciliği yöntemlerinin uygulanması
- Örüntü Değerlendirme: Bazı ölçütlere göre elde edilmiş ilginç örüntüleri bulmak ve değerlendirmek
- Bilgi Sunumu : Elde edilen bilgilerin kullanıcılara sunumunu

Veri Madenciliği Uygulama Alanları

Bilim	İş Hayatı	Web	Devlet
• Astronomi • Biyoinformatik • İlaç keşfi	• Reklam • CRM (Müşteri İlişkileri Yönetimi) ve Müşteri Modelleme • E-ticaret • Yatırım değerlendirme ve karşılaştırma • Sağlık • Üretim • Spor/eğlence • Telekom (telefon ve iletişim) • Hedef pazarlama	• Metin Madenciliği (haber grubu, email, dokümanlar) • Web analizi • Arama motorları	• Terörle Mücadele • Kanun Yaptırımı • Vergi Kaçakçılarının Profilinin Çıkarılması

Uygulamalar

- Hangi promosyonu ne zaman uygulamalıyım?
- Hangi müşteri aldığı krediyi geri ödemeyebilir?
- Bir müşteriye ne kadar kredi verilebilir?
- Sahtekarlık olabilecek davranışlar hangileridir?
- Hangi müşteriler yakın zamanda kaybedilebilir?
- Hangi müşterilere promosyon yapmalıyım?
- Hangi yatırım araçlarına yatırım yapmalıyım?

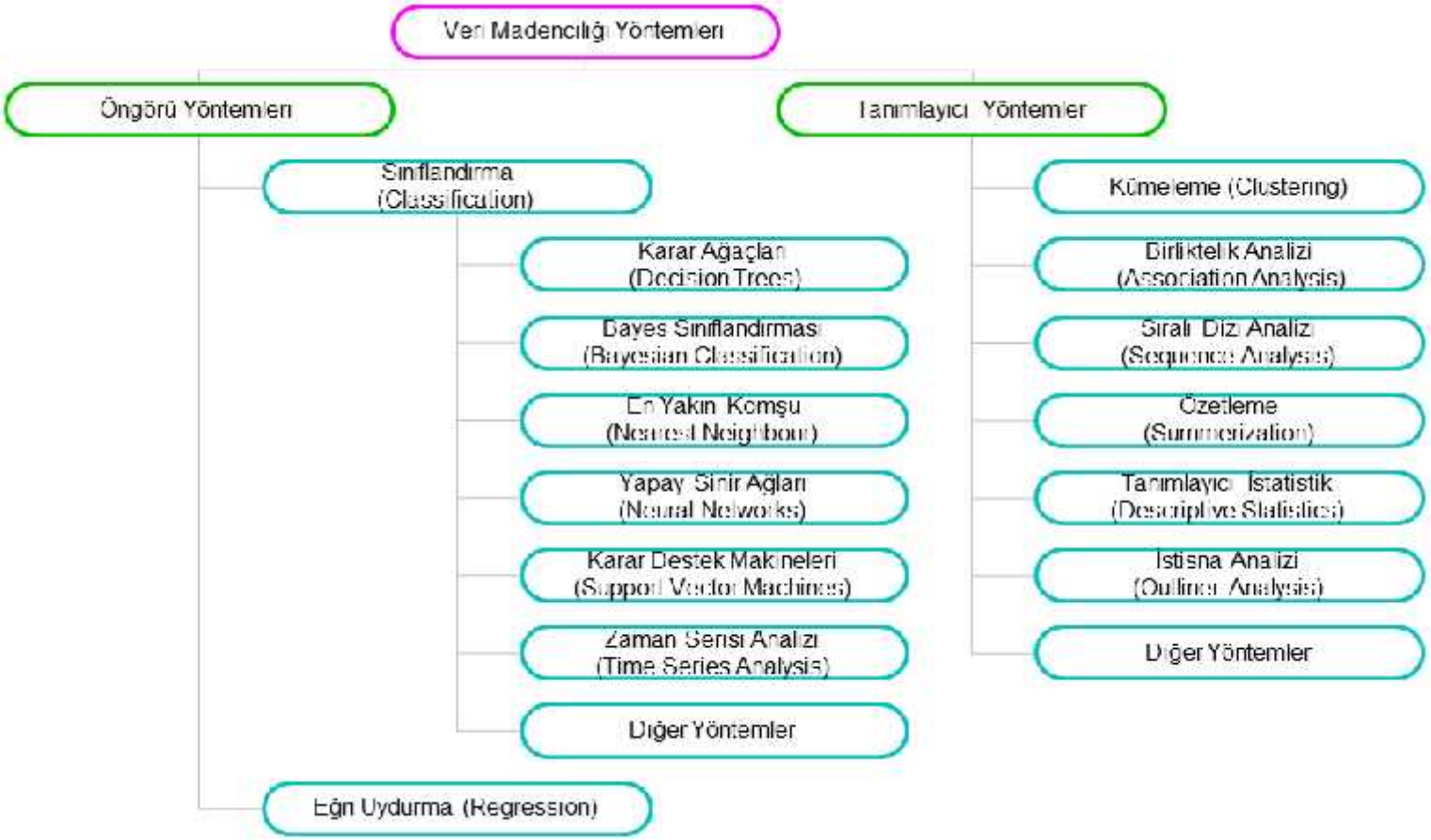
Veri Kaynakları

- Veri dosyaları
- Veritabanı kaynaklı veri kümeleri
  - ilişkisel veritabanları, veri ambarları
- Gelişmiş veri kümeleri
  - duraksız veri (data stream), algılayıcı verileri (sensor data)
  - zaman serileri, sıralı diziler (biyolojik veriler)
  - çizgeler, sosyal ağ (social networks) verileri
  - konumsal veriler (spatial data)
  - çoğul ortam veritabanları (multimedia databases)
  - nesneye dayalı veritabanları
  - www

Veri Madenciliği Yöntemleri

- amaç: veriyi belli bir modele uydurmak
  - tanımlayıcı
    - En iyi müşterilerim kimler?
    - Hangi ürünler birlikte satılıyor?
    - Hangi müşteri gruplarının alışveriş alışkanlıkları benzer?
  - kestirime dayalı
    - Kredi başvuralarını risk gruplarına ayırma
    - Şirketle çalışmayı bırakacak müşterileri öngörme
    - Borsa tahmini

Veri Madenciliği Yöntemleri



Veri Madenciliği İşlevleri (1/2)

- Sınıflandırma (Classification): Veriyi önceden belirlenmiş sınıflardan birine dahil eder.
  - Danışmanlı (Gözetimli) öğrenme
  - Örüntü tanıma
- Eğri uydurma (Regression): Veriyi gerçel değerli bir fonksiyona dönüştürür.
- Zaman serileri inceleme (Time Series Analysis): Zaman içinde değişen verinin değerini öngörür.
- İstisna Analizi (Outlier Analysis): Verinin geneline uymayan nesneleri belirleme

Veri Madenciliği İşlevleri (2/2)

- Kümeleme (Clustering): Benzer verileri aynı grupta toplama
  - Danışmansız (Gözetimsiz) öğrenme
- Özetleme (Summarization): Veriyi alt gruplara ayırır. Her alt grubu temsil edecek özellikler bulur.
- İlişkilendirme kuralları (Association Rules)
  - Veriler arasındaki ilişkiyi belirler
- Sıralı dizileri bulma (Sequence Discovery): Veri içinde sıralı örüntüler bulmak için kullanılır.

Veri Madenciliğinde Sorunlar (1/3)

- Gizlilik ve sosyal haklar
  - Kişilere ait verilerin toplanarak, kişilerden habersiz ve izinsiz olarak kullanılması
  - Veri madenciliği yöntemleri ile bulunan sonuçların izinsiz olarak açıklanması (/paylaşılması)
  - Gizlilik ve veri madenciliği politikalarının düzenlenmesi
- Kullanıcı Arabirimi
  - Görüntüleme
    - Sonucun anlaşılabilir ve yorumlanabilir hale getirilmesi
    - Bilginin sunulması
  - Etkileşim
    - Veri madenciliği ile elde edilen bilginin kullanılması
    - Veri madenciliği yöntemine müdahale etmek
    - Veri madenciliği yönteminin sonucuna müdahale etmek
- Veri madenciliği yöntemi
- Başarım ve ölçeklenebilirlik

Veri Madenciliğinde Sorunlar (2/3)

- Veri madenciliği yöntemi
  - Farklı tipte veriler üzerinde çalışabilme
  - Farklı seviyelerde kullanıcı ile etkileşim halinde olabilme
  - Uygulama ortamı bilgisini kullanabilme
  - Veri madenciliği ile elde edilen sonucu anlaşılır şekilde sunabilme
  - Gürültülü ve eksik veri ile çalışabilme (ve iyi sonuç verebilme)
  - Değişen veya eklenen verileri kolayca kullanabilme
  - Örüntü değerlendirme: önemli örüntüleri bulma

Veri Madenciliğinde Sorunlar (3/3)

- Başarım ve ölçeklenebilirlik
  - Kullanabilirlik ve ölçeklenebilirlik
    - Zaman karmaşıklığı ve yer karmaşıklığı kabul edilebilir
    - Örnekleme yapabilme
  - Paralel ve dağıtık yöntemler
    - Artımlı veri madenciliği
    - Parçala ve çöz

VERİ **ÖNİŞLEME**

(Veri Önişleme-1)

Veri Önişleme

- Veri
- Veri Önişleme
  - Veriyi Tanıma
  - Veri temizleme
  - Veri birleştirme
  - Veri dönüşümü
  - Veri azaltma
- Benzerlik ve farklılık

Veri Nedir?

- Nesneler ve nesnelerin niteliklerinden oluşan küme
  - kayıt (record), varlık (entity), örnek (sample, instance) nesne için kullanılabilir.
- Nitelik (attribute) bir nesnenin (object) bir özelliğidir
  - bir insanın yaşı, ortamın sıcaklığı...
  - boyut (dimension), özellik (feature, characteristic) olarak da kullanılır.
- Nitelikler ve bu niteliklere ait değerler bir nesneyi oluşturur.

Tid	Geri Ödeme	Medeni Durum	Gelir	Dolandırıcı
1	Evet	Bekar	125K	-1
2	Hayır	Evli	100K	-1
3	Hayır	Bekar	70K	-1
4	Evet	Evli	120K	-1
5	Hayır	Boşanmış	95K	1
6	Hayır	Evli	60K	-1
7	Evet	Boşanmış	220K	-1
8	Hayır	Bekar	85K	1
9	Hayır	Evli	75K	-1
10	Hayır	Bekar	90K	1

Değer Kümeleri

- Nitelik için saptanmış sayılar veya semboller
- Nitelik & Değer Kümeleri
  - aynı nitelik farklı değer kümelerinden değer alabilir
    - ağırlık: kg, lb(libre, ağırlık ölçüsü)
  - farklı nitelikler aynı değer kümesinden değer alabilirler
    - ID, yaş: her ikisi de sayısal

İstatistiksel Veri Türleri

- 1- Nümerik Veriler : Sayısal-Nümerik-Nicel Veriler de denmektedir. Boy,Yaş gibi süreklilik arzeden değerler Nümerik verilerdir. “Daha fazla” ifadesi ile kullanılabilirler. Sürekli ve süreksiz olarak iki başlıkta ele alınabilir:
  - a) Sürekli Nümerik Veriler: Yaş, Sıcaklık
  - b) Aralıklı Nümerik Veriler (Interval): Çocuk Sayısı, Kaza Sayısı

- 2-Nominal Veriler : Kategorik bir veri çeşididir.“Daha fazla” ifadesi ile kullanılmazlar. İkiye ayrılır:
  - a)Binary Veriler: Var-Yok, Kadın-Erkek, Hasta-Sağlıklı
  - b)İkiden Çok Kategorili: Medeni Durum-Renk-Irk-Şehir, İsim, Forma Numarası
  - Örneğin forma numarası oyuncunun seviyesi ile ilgili bir bilgi içermez.

İstatistiksel Veri Türleri

- 3-Ordinal Veriler : Ordinal veriler de yine kategorik veri türündendir. Fakat değerleri arasında sıralı bir ilişki bulunmaktadır. “Daha fazla” ifadesi ile kullanılabilirler ancak nekadar daha fazla olduğunun ölçüsünü veremezler. Örneğim: Eğitim Düzeyi, Sosyoekonomik ölçek skorları gibi. Nominal veriler, ordinal verilere göre daha az bilgi taşırlar.

- 4-Ratio Veriler : Nümerik verilere benzerler. 100 santigrat derece, 50 santrigat derecenin iki katı denilemez ama derece kelvine çevrilirse 60 kelvin 30 kelvinin 2 misli sıcak denilebilir. Oran verilebilir veri türlerine Ratio veriler denir. Burada kelvin derece ratio türünden bir değişken iken, santigrat ise nümerik veri türüne örnek olarak verilebilir.

Nitelik Türleri

- Belli aralıkta yeralan değişkenler (interval)
  - sıcaklık, tarih
- İkili değişkenler (binary)
  - cinsiyet
- Ayrık ve sıralı değişkenler
  - göz rengi, posta kodu

Problem

- Gerçek uygulamalarda toplanan veri kirli
  - eksik: bazı nitelik değerleri bazı nesneler için girilmemiş, veri madenciliği uygulaması için gerekli bir nitelik kaydedilmemiş
    - meslek = “ ”
  - gürültülü: hatalar var
    - maaş= “-10”
  - tutarsız: nitelik değerleri veya nitelik isimleri uyumsuz
    - yaş= “35”, d.tarihi: “03/10/2004”
    - önceki oylama değerleri: “1,2,3”, yeni oylama değerleri: “A,B,C”
    - bir kaynakta nitelik değeri ‘ad’, diğerinde ‘isim’

Veri kirliliği örneği-1

kapsam	sorun	Kirli veriler	sebep
özellik	Yanlış değer	Doğum_günü =30.13.1990	Değerler alan dışındadır
Kayıt	Özellikler arasında bağımlılığın yanlış olması	Yaş=42 Doğum_günü=12.02.1990	«yaş»la doğum günü değerleri tutarsızdır
Kayıt türü	Eşsizliğin bozulması	Pers1=(ad=«Ali Yavuz», pno=«123456» Pers1=(ad=«Metin SAĞLAM», pno=«123456»	Personel numarasının eşsiz olması koşulu bozulmuştur
kaynak	Erişimsel bütünlüğün bozulması	Pers1=(ad=«Metin SAĞLAM», şube_no=«123456»	«123456»no'lu şube tanımlanmamıştır

Veri kirliliği örneği

kapsam	sorun	Kirli veriler	sebep
özellik	Değer yoktur	Tel:=285218 163	Rakam eksiktir
özellik	Kelimenin yanlış yazılışı	Kent=«Trabzun»	Fonetik hata
özellik	yanlış alan değeri	Kent=«İtalya»	«Italya» «kent» alanına dahil değil
kayıt	Özellikler arası bağımlılığın bozulması	Kent=«Çanakkale»; plaka_no=19	«Çanakkale'nin plaka numarası 19 değil
Kayıt türü	Kelimelerin farklı dizilişi	Ad1 =«Kerim UĞUR» Ad2=«YILMAZ Temel»	Ad ve soyadların sıraları farklıdır
Kayıt türü	Kayıtlarda zıtlık	Pers1=(ad=«Ali Yavuz», doğum_tar=12.12.1995 Pers2=(ad=«Ali Yavuz», doğum_tar=10.09.1995	Aynı varlık farklı değerlerle tanımlanmıştır

Verinin Kirli Olma Nedenleri

- Eksik veri kayıtlarının nedenleri
  - Veri toplandığı sırada bir nitelik değerinin elde edilememesi, bilinmemesi
  - Veri toplandığı sırada bazı niteliklerin gerekliliğinin görülememesi
  - İnsan, yazılım ya da donanım problemleri
- Gürültülü (hatalı) veri kayıtlarının nedenleri
  - Hatalı veri toplama gereçleri
  - İnsan, yazılım ya da donanım problemleri
  - Veri iletimi sırasında problemler
- Tutarsız veri kayıtlarının nedenleri
  - Verinin farklı veri kaynaklarında tutulması
  - İşlevsel bağımlılık kurallarına uyulmaması

Sonuç

- Veri güvenilmez olduğunda:
  - Veri madenciliği sonuçlarına güvenilebilir mi?
  - Kullanılabilir veri madenciliği sonuçları kaliteli veri ile elde edilebilir.
- Veri kaliteli ise veri madenciliği uygulamaları ile yararlı bilgi bulma şansı daha fazla.

Veriyi Tanıma

Veriyi Tanımlayıcı Özellikler

- Amaç: Veriyi daha iyi anlamak
  - Merkezi eğilim (central tendency), varyasyon, yayılma, dağılım
- Verinin dağılım özellikleri
  - Ortanca, en büyük, en küçük, sıklık derecesi, aykırılık, varyans
- Sayısal nitelikler -> sıralanabilir değerler
  - verinin dağılımı
  - kutu grafiği çizimi ve sıklık derecesi incelemesi

Veri Önişleme

- Veri temizleme
  - Eksik nitelik değerlerini tamamlama, hatalı veriyi düzeltme, aykırılıkları saptama ve temizleme, tutarsızlıkları giderme
- Veri birleştirme
  - Farklı veri kaynağındaki verileri birleştirme
- Veri dönüşümü
  - Normalizasyon
- Veri azaltma
  - Aynı veri madenciliği sonuçları elde edilecek şekilde veri miktarını azaltma

Veri Temizleme

- Gerçek uygulamalarda veri eksik, gürültülü veya tutarsız olabilir.
- Veri temizleme işlemleri
  - Eksik nitelik değerlerini tamamlama
  - Aykırılıkların bulunması ve gürültülü verinin düzeltilmesi
  - Tutarsızlıkların giderilmesi

Eksik Veri

- Veri için bazı niteliklerin değerleri her zaman bilinemeyebilir.
- Eksik veri
  - diğer veri kayıtlarıyla tutarsızlığı nedeniyle silinmesi
  - bazı nitelik değerleri hatalı olması dolayısıyla silinmesi
  - yanlış anlama sonucu kaydedilmeme
  - veri girişi sırasında bazı nitelikleri önemsiz görme

Eksik Veriler nasıl Tamamlanır?

- Eksik nitelik değerleri olan veri kayıtlarını kullanma (sil)
- Eksik nitelik değerlerini elle doldur
- Eksik nitelik değerleri için global bir değişken kullan (Null, bilinmiyor,...)
- Eksik nitelik değerlerini o niteliğin ortalama değeri ile doldur
- Aynı sınıfa ait kayıtların nitelik değerlerinin ortalaması ile doldur
- Olasılığı en fazla olan nitelik değeriyle doldur

Age	Income	Team	Gender
23	24,200	Red Sox	M
39	?	Yankees	F
45	45,390	?	F

Olasılığı en fazla olan nitelik değeriyle doldur

Eksik nitelik değerlerini o niteliğin ortalama değeri ile doldur

Aynı sınıfa ait kayıtların nitelik değerlerinin ortalaması ile doldur

Gürültülü Veri

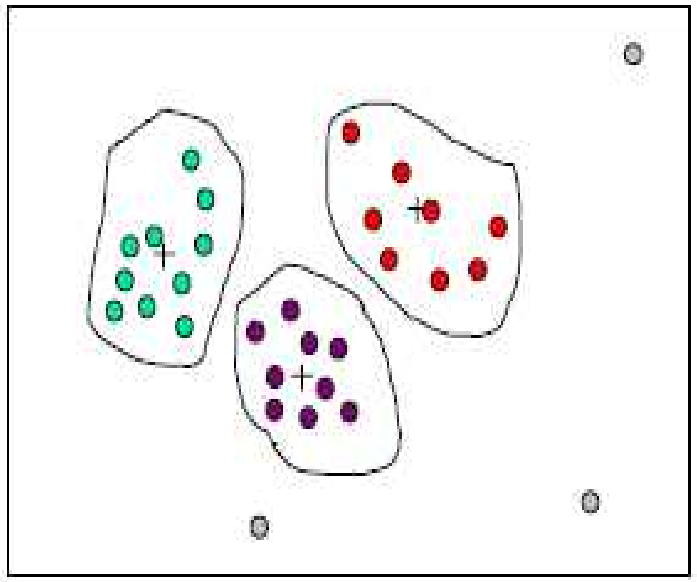
- Ölçülen bir değerdeki hata
- Yanlış nitelik değerleri
  - hatalı veri toplama gereçleri
  - veri girişi problemleri
  - veri iletimi problemleri
  - teknolojik kısıtlar
  - nitelik isimlerinde tutarsızlık

Gürültülü Veri nasıl düzeltilir?

- Gürültüyü yok etme
  - Kümeleme
  - Eğri uydurma
    - veriyi bir fonksiyona uydurarak gürültüyü düzeltir.

Kümeleme

- Benzer veriler aynı kümede olacak şekilde gruplanır
- Bu kümelerin dışında kalan veriler aykırılık olarak belirlenir ve silinir.



Eğri Uydurma

- Veri bir fonksiyona uydurulur. Doğrusal eğri uydurmada, bir değişkenin değeri diğer bir değişken kullanılarak bulunabilir.

Veri Birleştirme

Veri Birleştirme

- Farklı kaynaklardan verilerin tutarlı olarak birleştirilmesi
- Şema birleştirilmesi
  - Aynı varlıkların saptanması
- Nitelik değerlerinin tutarsızlığının saptanması
  - Aynı nitelik için farklı kaynaklarda farklı değerler olması
  - Farklı metrikler kullanılması

Gereksiz Veri

Farklı veri kaynaklarından veriler birleştirilince gereksiz (fazla) veri oluşabilir

- aynı nitelik farklı kaynaklarda farklı isimle
- bir niteliğin değeri başka bir nitelik kullanılarak hesaplanabilir

VERİ MADENCİLİĞİ

(Veri Ön İşleme-2)

Veri Dönüşümü

- Veri, veri madenciliği uygulamaları için uygun olmayabilir
- Seçilen algoritmaya uygun olmayabilir
  - Veri belirleyici değil

Çözüm

- Veri düzeltme
- Normalizasyon

Normalizasyon

- min-max normalizasyon
  - min-max normalleştirmesi ile orijinal veriler yeni veri aralığına doğrusal dönüşüm ile dönüştürülürler. Bu veri aralığı genellikle 0-1 aralığıdır.
- z-score normalizasyon
  - z Skor normalleştirmede (veya 0 ortalama normalleştirme) ise değişkenin her hangi bir y değeri, değişkenin ortalaması ve standart sapmasına bağlı olarak bilinen Z dönüşümü ile normalleştirilir.
- ondalık normalizasyon
  - Ondalık ölçekleme ile normalleştirmede ise, ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme gerçekleştirilir. Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Ondalık ölçeklemenin formülü aşağıdaki şekildedir:
  - Örneğin 900 maksimum değer ise, n=3 olacağından 900 sayısı 0,9 olarak normalleştirilir.

Normalizasyon

- min-max normalizasyon

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalizasyon

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- ondalık normalizasyon

$$v' = \frac{v}{10^j}$$

j: Max($|v'|$)<1 olacak şekildeki en küçük tam sayı

Normalizasyon

- **Min-max normalizasyon**:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ör. Yıllık gelir $12,000 ile $98,000 arasını [0.0, 1.0] aralığına normalize edelim. $73,000 kaça denk gelir? $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$
- **Z-score normalizasyon** (μ: ortalama, σ: standard sapma):

$$v' = \frac{v - «_A}{\pm_A}$$

  - Ör. Let μ = 54,000, σ = 16,000. Öyleyse: $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$
- **Ondalıklı Normalizasyon**

$$v' = \frac{v}{10^j}$$

$73,000 kaça denk gelir? v'= 0.73

Nitelik Oluşturma

- Yeni nitelikler yarat
  - orjinal niteliklerden daha önemli bilgi içersin
    - alan=boy x en
  - veri madenciliği algoritmalarının başarımı daha iyi olsun

Veri Azaltma

Veri Azaltma

- Veri miktarı çok fazla olduğu zaman veri madenciliği algoritmalarının çalışması ve sonuç üretmesi çok uzun sürebilir
  - veriyi azaltma başarımı artırır
  - sonucun (nerdeyse) hiç değişmemesi gerekir
- Veri azaltma
  - nitelik azaltma
  - veri sıkıştırma
  - veri ayrıklaştırma
  - veri küçültme

Nitelik Azaltma

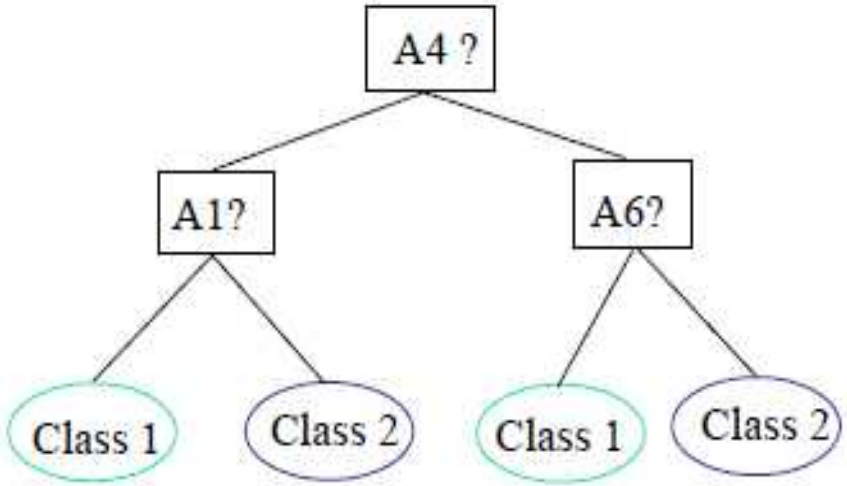
- Nitelikler kümesinin bir alt kümesi seçilerek veri madenciliği işlemi yapılır.
- d boyutlu veri kümesi k<d olacak şekilde k boyuta taşınır.

- Nitelik seçme
  - Veri madenciliği uygulaması için gerekli olan niteliklerin seçilmesi

Örnek

Başlangıç nitelikler kümesi:

{A1, A2, A3, A4, A5, A6}

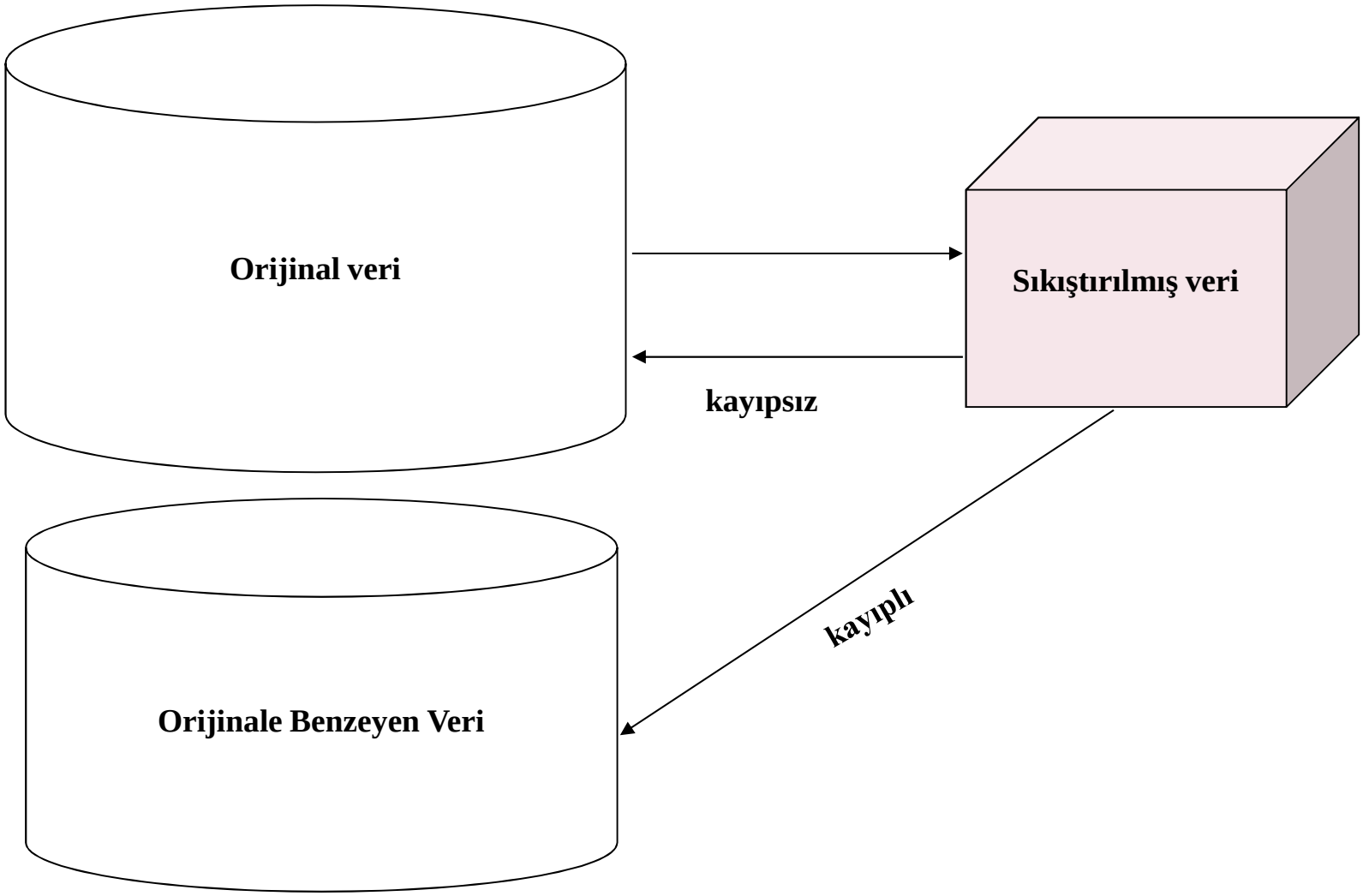


Seçilen nitelik kümesi: {A1, A4, A6}

Veri Sıkıştırma

- Verinin boyutunu azaltır
  - daha az saklama ortamı
  - veriye ulaşmak daha çabuk
- Kayıplı ve kayıpsız veri sıkıştırma
  - bazı yöntemler bazı veri tiplerine uygun
- Eğer veri madenciliği yöntemi sıkıştırılmış veri üzerinde doğrudan çalışabiliyorsa elverişli

Veri Sıkıştırma

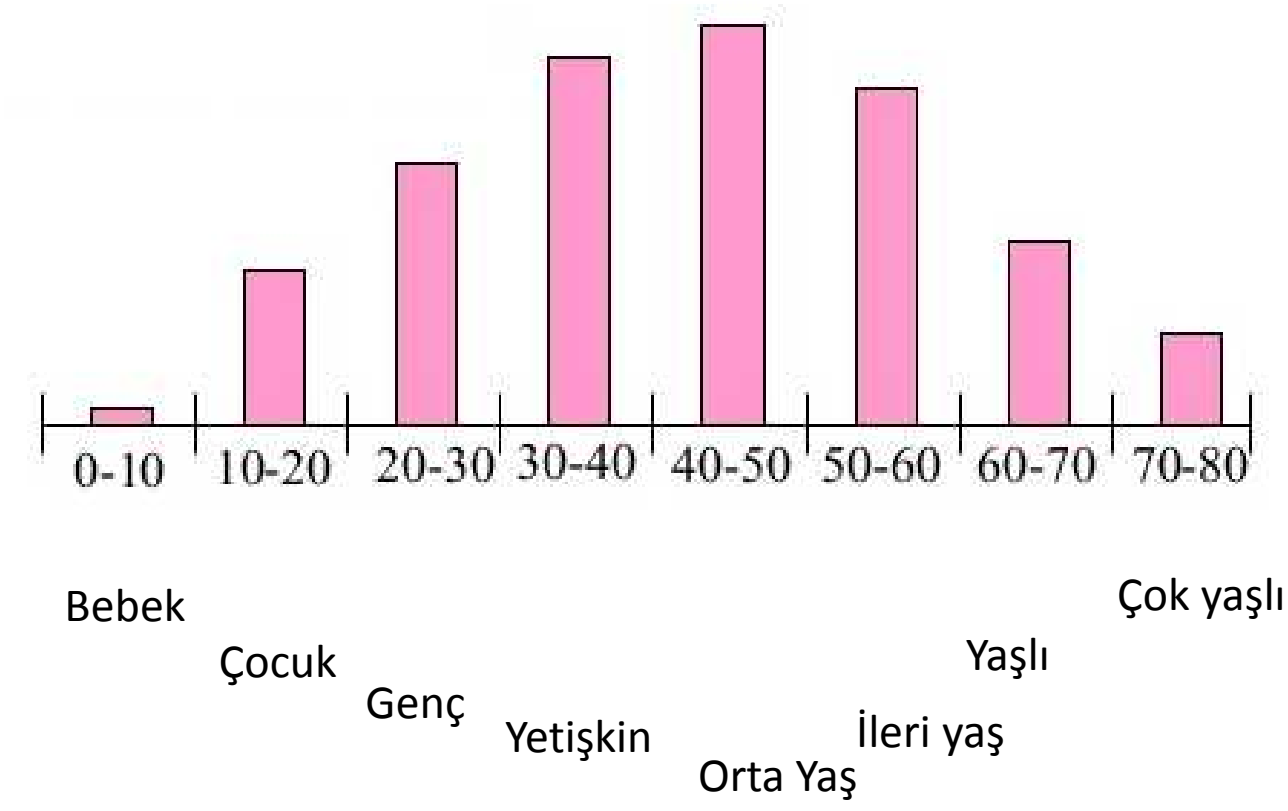


Veri Ayrıklaştırma

- Bazı veri madenciliği algoritmaları sadece ayrık veriler ile çalışır.
- Sürekli bir nitelik değerini bölerek her aralığı etiketler.
- Verinin değeri, bulunduğu aralığın etiketi ile değişir.
- Veri boyutu küçülür.

Veri Ayrıklaştırma

Müşteri Yaşına göre ayrıklaştırma

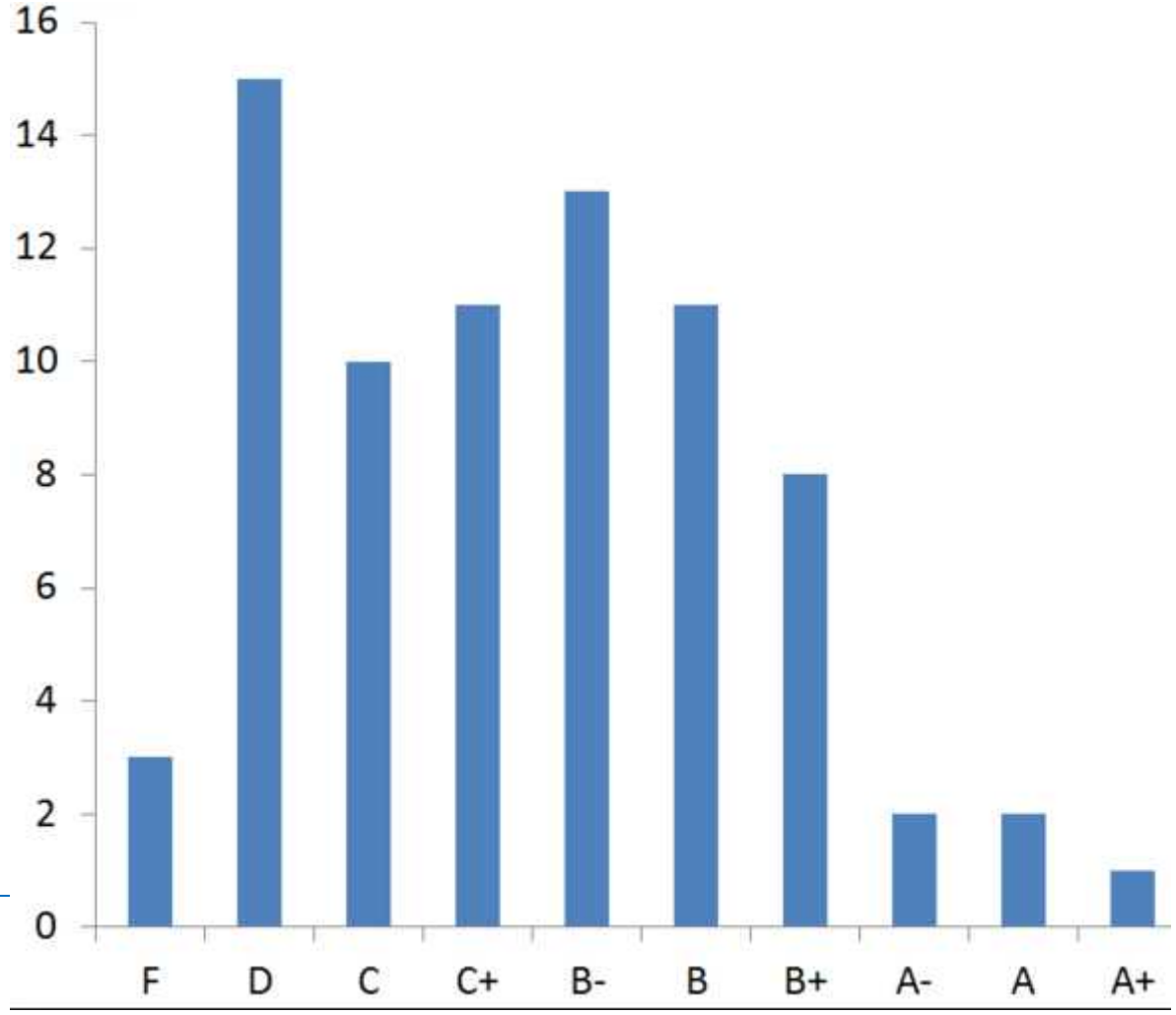


Veri Küçültme

- Veriyi farklı şekillerde gösterme
  - histogram
  - kümeleme
  - örnekleme

Histogram ile Veri Küçültme

- Verinin dağılımı
- Veriyi bölerek her bölüm için veri değerini gösterir (toplam, ortalama)

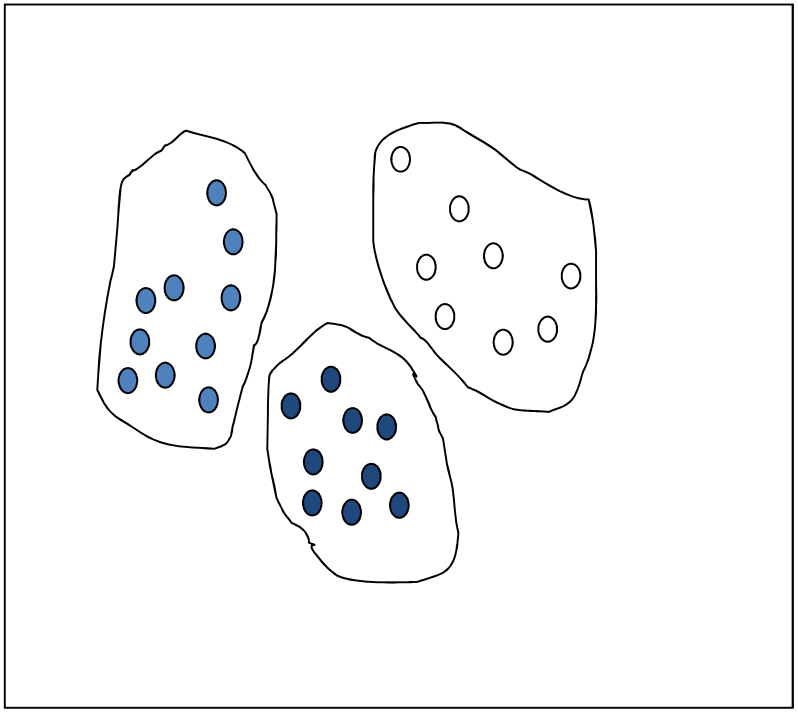


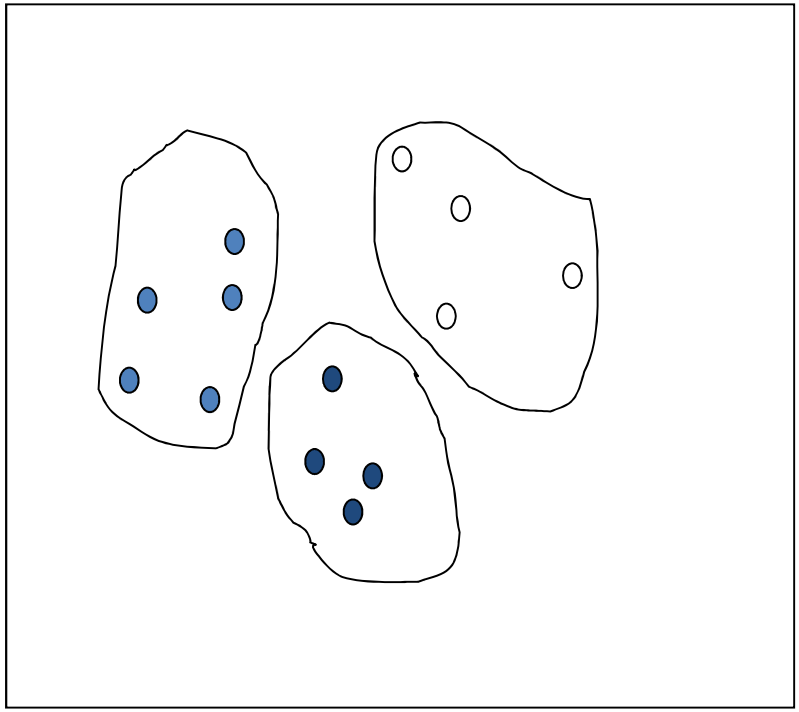
Kümeleme ile Veri Küçültme

- Veri kümelere ayrılır
- Veri kümeleri temsil eden örnekler (küme merkezleri) ve aykırılıklar ile temsil edilir
- Etkisi verinin dağılımına bağlı.

Kümeleme ile veri küçültme

- Kümelenmiş veri
- Her kümeden orantılı sayıda temsilci seçimi

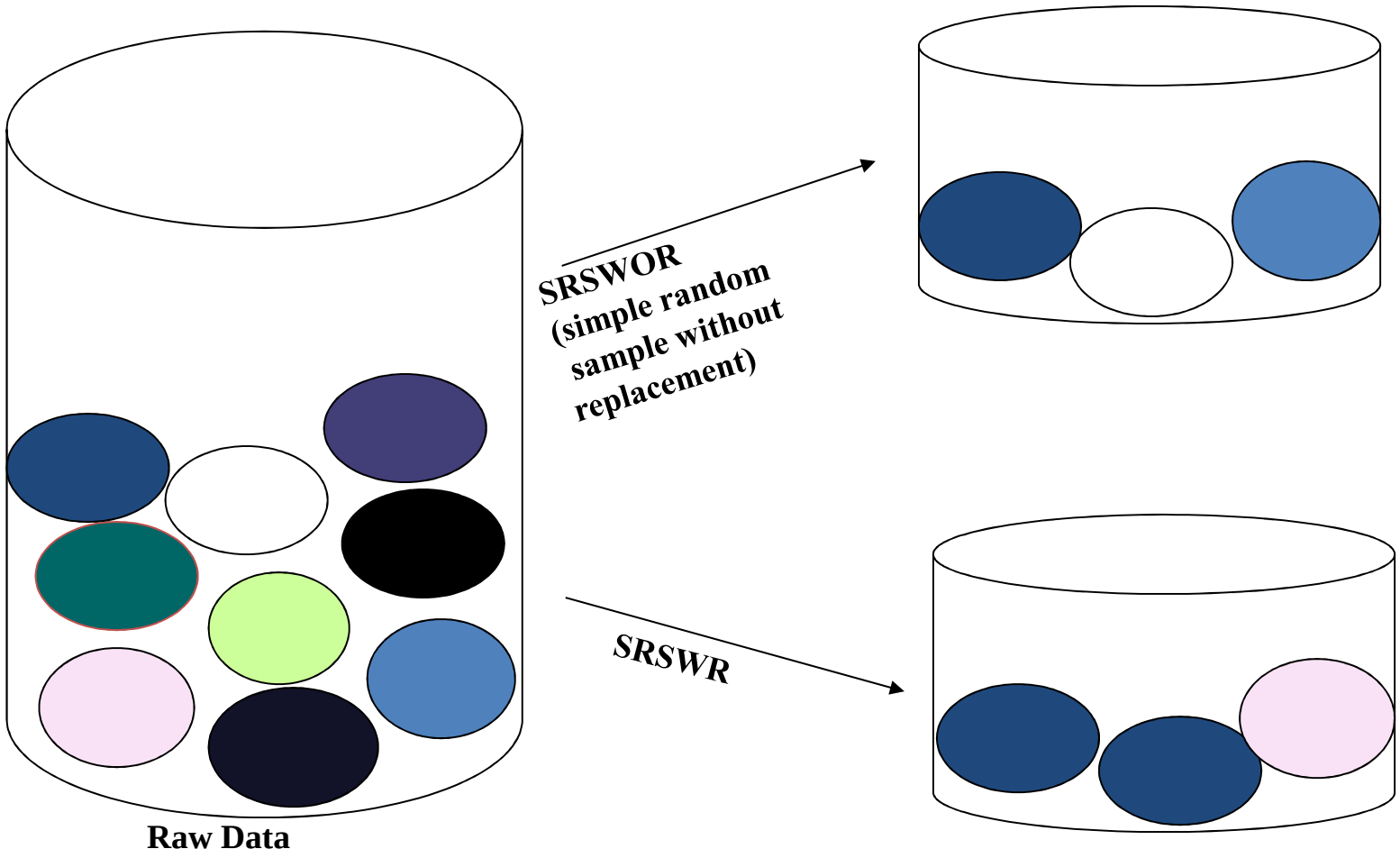




Örnekleme ile Veri Küçültme

- Büyük veri kümesini daha küçük bir alt küme ile temsil etme
- Alt küme nasıl seçiliyor?
  - yerine koymadan örnekleme (SRSWOR)
  - yerine koyarak örnekleme (SRSWR)
  - katman örnekleme (katman: nitelik değerine göre grup)

Örnekleme


SRSWOR
(simple random
sample without
replacement)
SRSWR
Raw Data

Benzerlik ve Farklılık

Benzerlik ve Farklılık

- Benzerlik
  - iki nesnenin benzerliğini ölçen sayısal değer
  - nesneler birbirine daha benzer ise daha büyük
  - genelde 0-1 aralığında değer alır
- Farklılık
  - iki nesnenin birbirinden ne kadar farklı olduğunu gösteren sayısal değer
  - nesneler birbirine daha benzer ise daha küçük
  - en küçük farklılık genelde 0
  - üst sınır değişebilir.

Uzaklık Çeşitleri

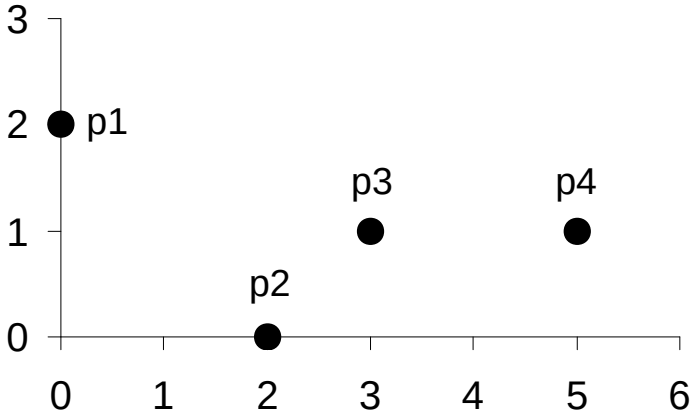
- Öklid
- Minkowski (Manhattan)

Öklid Uzaklığı

- Öklid uzaklığı (Euclidean Distance) nesneler arası farklılığı bulmak için kullanılır.
  - *p* adet niteliği (boyutu) olan *i* ve *j* nesneleri arasındaki uzaklık

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2}$$

Öklid Uzaklığı



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Uzaklık Matrisi

Minkowski Uzaklığı

- Öklid uzaklığının genelleştirilmiş hali

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+...+|x_{ip}-x_{jp}|^q)}$$

q: positif tam sayı

- q=1 → Manhattan uzaklığı

Minkowski Uzaklığı

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Manhattan Uzaklık Matrisi

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

Öklid Uzaklık Matrisi

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Benzerlik Özellikleri

- İki nesne arası benzerlik özellikleri
- 1. sim(i,j)>=0
- 2. sim(i,j)=sim(j,i)

İkili Değişkenler Arası Benzerlik

- İkili bir değişkenin 0 veya 1 olarak iki değeri olabilir.
- Bir olasılık tablosu oluşturulur:

		Nesne j	
		0	1
Nesne i	0	M_{00}	M_{01}
	1	M_{10}	M_{11}

M_{00}: i nesnesinin 0, j nesnesinin 0 olduğu niteliklerin sayısı
M_{10}: i nesnesinin 1, j nesnesinin 0 olduğu niteliklerin sayısı
M_{01}: i nesnesinin 0, j nesnesinin 1 olduğu niteliklerin sayısı
M_{11}: i nesnesinin 1, j nesnesinin 1 olduğu niteliklerin sayısı

- Yalın uyum katsayısı (simple matching coefficient): ikili değişkenin simetrik olduğu durumlarda

$$sim(i,j) = \frac{M_{11}+M_{00}}{M_{00}+M_{01}+M_{10}+M_{11}}$$

- Jaccard katsayısı (İkili değişkenin asimetrik olduğu durumlar):

$$d(i,j) = \frac{M_{11}}{M_{01}+M_{10}+M_{11}}$$

Örnek

- Jaccard

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	1	0	1	0	0	0
Mary	1	0	1	0	1	0
Jim	1	1	0	0	0	0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$